**GRoup - 09**

**Subhajit Chowdhury**

**Ashutosh Jha**

**Suvendu Kumar Lenka**

**Sathyanarayana M N**

**Sowjanya Tallavajuala**

**Tarun Kamal Khian**

Big Data Analytics

Assignment – Predict house prices

Table of Contents

[Goal 3](#_Toc490390884)

[Solution 3](#_Toc490390885)

[Exploratory Data Analysis 3](#_Toc490390886)

[Identify the important predictor variables 4](#_Toc490390887)

[Predictive model 5](#_Toc490390888)

[Predictive PErformance 7](#_Toc490390889)

[Conclusion 8](#_Toc490390890)

# Goal

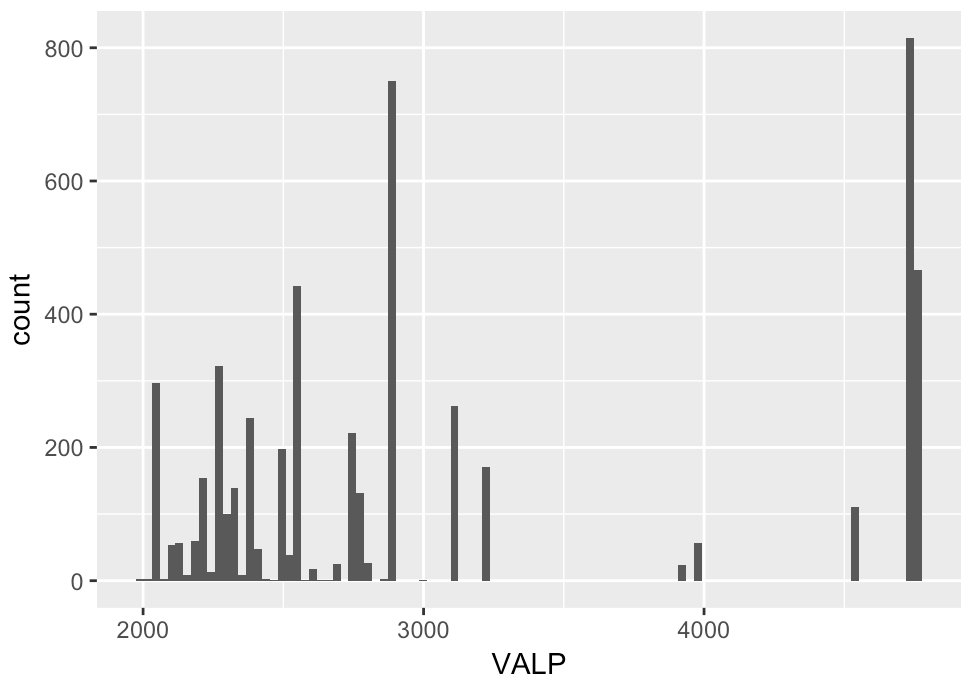
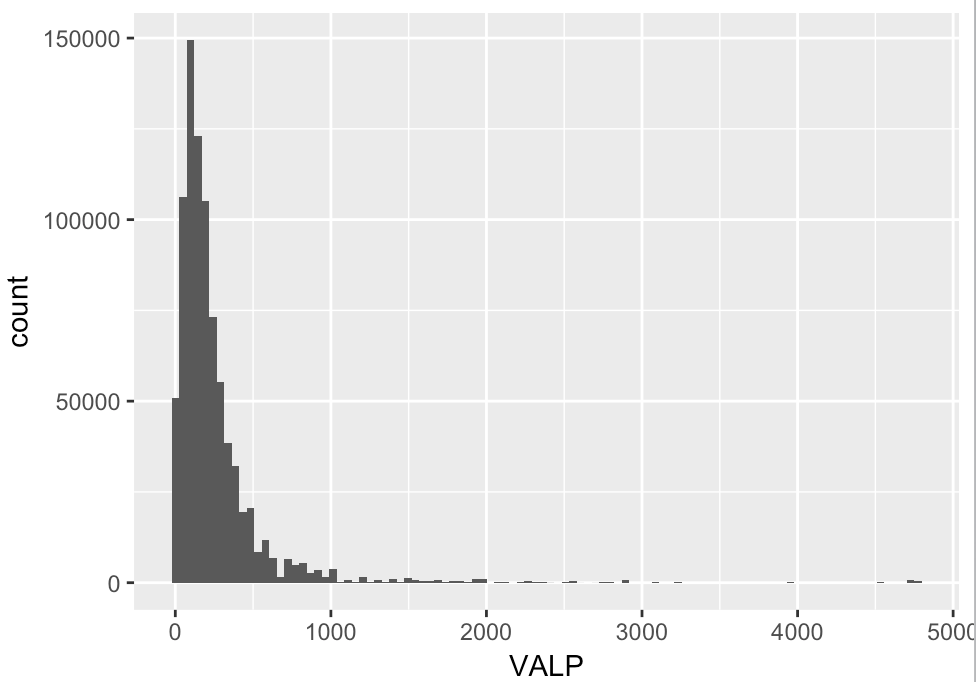
The goal of this assignment is to predict property value in relation to other variables in the 2013 American Community Survey.

# Solution

## Exploratory Data Analysis

Important Observations:

* Replicate weights can be useful to understand the range (confidence interval) of a specific value but we decided to omit it. For more details visit: <https://usa.ipums.org/usa/repwt.shtml>
* Allocations where used to fill values where it was not available. Hence does not make sense for us to include it into our dataset. For more details visit: <https://usa.ipums.org/usa/flags.shtml>
* Poverty Value is not available for group quarters hence our analysis/prediction is limited to Housing
* All numeric fields are scaled to represented in 1000’s
* NA’s are replaced by 0 for both factors and numeric fields. In some cases, like ELEP, GASP, FULP, WATP – NA, 1, 2 are replace by 0. This makes sense as we are treating NA is its own category and in case of numeric fields at 0 rather than a mean or median.
* The property value is highly skewed. Most of the house prices are less than 1 Million USD. Hence it might be good to understand or identify clusters that would differentiate property prices that are more than 2M and less than 2M.



## Identify the important predictor variables

We run a Random Forest Regression to identify the important variables. The first 52 variables are important. TAXP, SMPCP, MRGP, INSP, HINCP are most important and we have colour coded the important variables that can be used in the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Importance** | **Feature** | **Importance** | **Feature** | **Importance** | **Feature** |
| 9.43561E-06 | LAPTOP | 0.00838219 | MRGT | 0 | OTHSVCEX |
| 8.59532E-06 | WORKSTAT | 0.006194991 | MRGI | 0 | REFR |
| 7.54269E-05 | NPF | 0.006156829 | BDSP | 0 | RNTM |
| 7.39134E-06 | HFL | 0.00558911 | RMSP | 0 | RNTP |
| 7.2886E-05 | VEH | 0.005519861 | TEN | 0 | RWAT |
| 6.52292E-05 | R60 | 0.00547255 | SMX | 0 | STOV |
| 5.19103E-05 | R65 | 0.005091761 | MRGX | 0 | TEL |
| 4.88537E-05 | FULP | 0.004298928 | BLD | 0 | TOIL |
| 4.53626E-05 | WKEXREL | 0.001511164 | ELEP | 0 | FFINCP |
| 4.38275E-05 | WATP | 0.000956617 | OCPIP | 0 | FGRNTP |
| 3.58214E-05 | CONP | 0.000824802 | ACR | 0 | GRNTP |
| 3.16272E-05 | FPARC | 0.00075163 | FSMOCP | 0 | GRPIP |
| 2.88143E-05 | SMP | 0.000629919 | MHP | 0 | HUGCL |
| 2.83163E-05 | NR | 0.000443757 | AGS | 0 | HUPAC |
| 2.17714E-06 | FHINCP | 0.000394186 | MV | 0 | HUPAOC |
| 2.14032E-05 | HHT | 0.000241426 | GASP | 0 | HUPARC |
| 1.8742E-05 | NOC | 0.000211774 | BUS | 0 | KIT |
| 1.73248E-05 | HHL | 0.000187256 | YBL | 0 | LNGI |
| 1.66429E-06 | BATH | 0.000172828 | WGTP | 0 | MULTG |
| 1.30077E-05 | FES | 0.000155386 | WIF | 0 | NPP |
| 1.20191E-05 | PARTNER | 0.000108238 | SVAL | 0 | NRC |
| 0.407016752 | TAXP | 0 | NP | 0 | PLM |
| 0.176612515 | SMOCP | 0 | ACCESS | 0 | PSF |
| 0.094885758 | MRGP | 0 | BROADBND | 0 | R18 |
| 0.077069988 | INSP | 0 | COMPOTHX | 0 | RESMODE |
| 0.069444505 | HINCP | 0 | DIALUP | 0 | SRNT |
| 0.035487263 | DIVISION | 0 | DSL | 0 | SSMC |
| 0.030825844 | FINCP | 0 | FIBEROP | 0 | RWATPR |
| 0.027823487 | ST | 0 | FS | 0 | SATELLITE |
| 0.018422664 | REGION | 0 | HANDHELD | 0 | SINK |
| 0.008476186 | PUMA | 0 | MODEM |  |  |

## Predictive model

We explore both the options of running regression on 52 important variables and only on 6 variables that have an importance of greater than 2%.

|  |  |  |
| --- | --- | --- |
| **Regression Model** | **R-Squared** | **Root Mean Square Error** |
| Linear Regression on 52 variables | 0.5054 | 238.7 |
| Linear Regression on 6 variables | 0.4344 | 255.3 |
| Linear Regression on 46 variables after removing variables that have a significant p-value | 0.5054 | 238.7 |

Hence, we chose the model with 46 variables and decided to validate the model performance on the test data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** | **Significance** |
| (Intercept) | -100.8800 | 1.01E+01 | -10.0352 | 2.2E-16 | \*\*\* |
| LAPTOP | -4.1299 | 9.32E-01 | -4.4301 | 0.000009423 | \*\*\* |
| WORKSTAT | 1.0375 | 4.99E-01 | 2.0802 | 0.03751 | \* |
| NPF | -7.9193 | 4.69E-01 | -16.8862 | 2.2E-16 | \*\*\* |
| HFL | 4.1537 | 2.43E-01 | 17.1107 | 2.2E-16 | \*\*\* |
| VEH | -2.0554 | 3.80E-01 | -5.4109 | 6.275E-08 | \*\*\* |
| R60 | 5.7793 | 7.13E-01 | 8.1079 | 4.441E-16 | \*\*\* |
| R65 | 8.3048 | 7.60E-01 | 10.9333 | 2.2E-16 | \*\*\* |
| FULP | -42.9600 | 5.86E-01 | -73.3529 | 2.2E-16 | \*\*\* |
| WKEXREL | 4.4250 | 2.51E-01 | 17.6003 | 2.2E-16 | \*\*\* |
| WATP | -32.3990 | 6.68E-01 | -48.5296 | 2.2E-16 | \*\*\* |
| CONP | -514.3500 | 4.92E+00 | -104.5156 | 2.2E-16 | \*\*\* |
| FPARC | -2.0565 | 3.24E-01 | -6.3489 | 2.17E-10 | \*\*\* |
| SMP | -519.5700 | 2.93E+00 | -177.2734 | 2.2E-16 | \*\*\* |
| NR | -14.9030 | 1.36E+00 | -10.9196 | 2.2E-16 | \*\*\* |
| FHINCP | 6.5886 | 7.29E-01 | 9.0345 | 2.2E-16 | \*\*\* |
| NOC | 3.6746 | 6.15E-01 | 5.9779 | 2.261E-09 | \*\*\* |
| HHL | 7.6268 | 4.39E-01 | 17.3813 | 2.2E-16 | \*\*\* |
| BATH | 47.3020 | 6.75E+00 | 7.0113 | 2.364E-12 | \*\*\* |
| FES | -8.2812 | 1.06E+00 | -7.8188 | 5.329E-15 | \*\*\* |
| TAXP | 0.2646 | 2.48E-02 | 10.6675 | 2.2E-16 | \*\*\* |
| SMOCP | 560.5800 | 1.94E+00 | 288.7077 | 2.2E-16 | \*\*\* |
| MRGP | -427.4700 | 1.90E+00 | -224.4558 | 2.2E-16 | \*\*\* |
| INSP | 25.1580 | 4.28E-01 | 58.8136 | 2.2E-16 | \*\*\* |
| HINCP | 0.4251 | 1.08E-02 | 39.3793 | 2.2E-16 | \*\*\* |
| DIVISION | 13.6360 | 1.46E-01 | 93.3835 | 2.2E-16 | \*\*\* |
| FINCP | 0.0696 | 1.07E-02 | 6.5148 | 7.283E-11 | \*\*\* |
| ST | -1.2610 | 2.13E-02 | -59.2798 | 2.2E-16 | \*\*\* |
| PUMA | 0.0013 | 3.47E-05 | 37.5313 | 2.2E-16 | \*\*\* |
| MRGT | -86.8880 | 1.38E+00 | -62.7672 | 2.2E-16 | \*\*\* |
| MRGI | -18.2130 | 1.23E+00 | -14.8088 | 2.2E-16 | \*\*\* |
| BDSP | 8760.4000 | 4.11E+02 | 21.3405 | 2.2E-16 | \*\*\* |
| RMSP | 5711.6000 | 1.76E+02 | 32.5393 | 2.2E-16 | \*\*\* |
| TEN | -55.3180 | 3.45E+00 | -16.0265 | 2.2E-16 | \*\*\* |
| SMX | 11.7140 | 9.63E-01 | 12.1684 | 2.2E-16 | \*\*\* |
| BLD | 18.6610 | 4.43E-01 | 42.152 | 2.2E-16 | \*\*\* |
| ELEP | -505.1200 | 4.25E+00 | -118.9646 | 2.2E-16 | \*\*\* |
| OCPIP | 80.5260 | 1.99E+01 | 4.0518 | 0.00005084 | \*\*\* |
| ACR | 24.9330 | 8.77E-01 | 28.4339 | 2.2E-16 | \*\*\* |
| FSMOCP | 14.5670 | 7.10E-01 | 20.5069 | 2.2E-16 | \*\*\* |
| MHP | -46.2380 | 5.54E-01 | -83.4351 | 2.2E-16 | \*\*\* |
| AGS | 1.5624 | 6.36E-01 | 2.4557 | 0.01406 | \* |
| GASP | -439.3400 | 4.82E+00 | -91.103 | 2.2E-16 | \*\*\* |
| YBL | -3.0379 | 1.17E-01 | -26.0056 | 2.2E-16 | \*\*\* |
| WGTP | -0.0449 | 4.85E-03 | -9.2453 | 2.2E-16 | \*\*\* |
| WIF | -10.0580 | 6.44E-01 | -15.6071 | 2.2E-16 | \*\*\* |
| SVAL | -11.5570 | 1.02E+00 | -11.3364 | 2.2E-16 | \*\*\* |

## Predictive PErformance

|  |  |
| --- | --- |
| Sample | Root Mean Square Error |
| Train Sample | 238.7 |
| Test Sample | 235.1341 |
| Test Sample where house prices > 2M USD | 2125.36 |
| Test Sample where house prices < 2M USD | 152.0634 |
| Test Sample where house prices < 1.3M USD | 129.0808 |

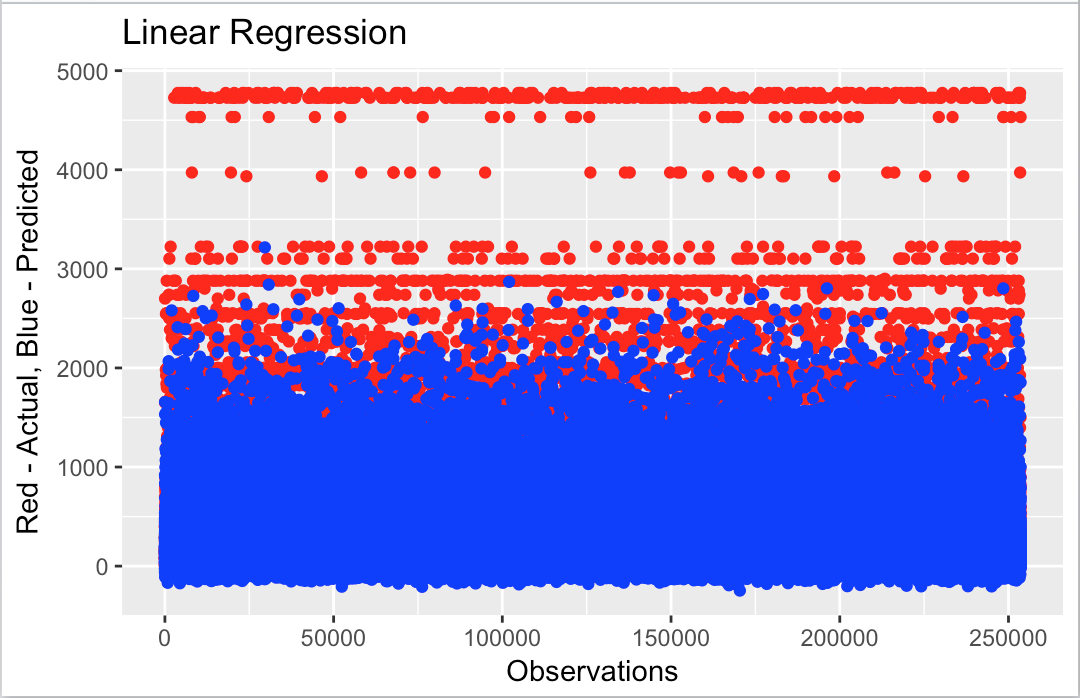
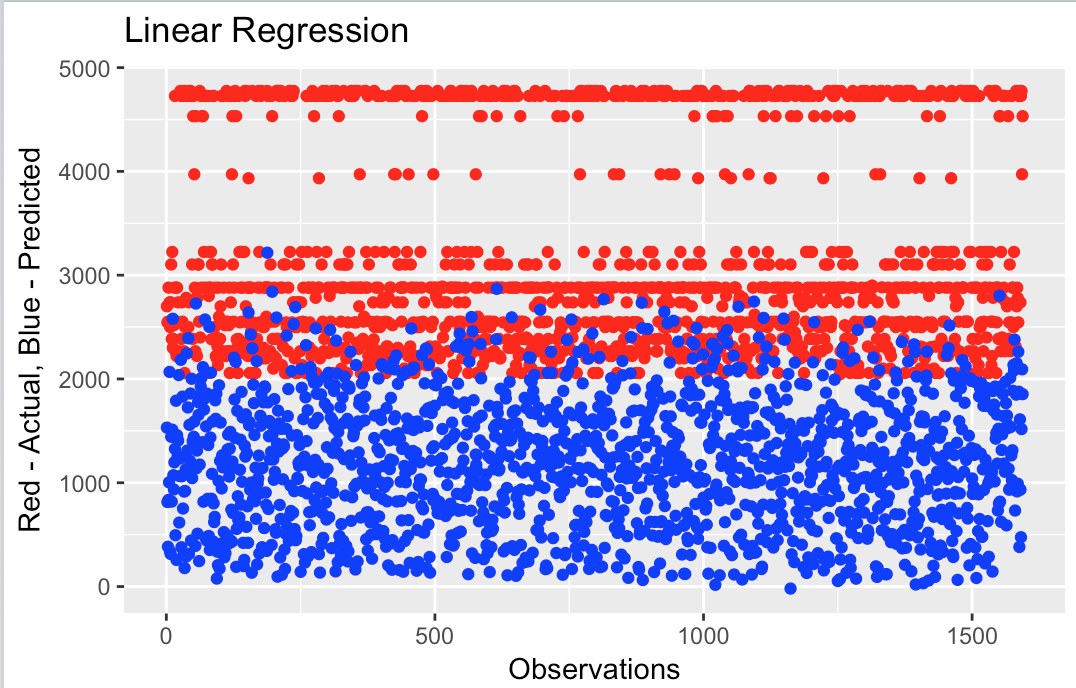
****

Figure 1: Actual vs Prediction >2M USD Figure 2: Actual vs Prediction

# Conclusion

* The linear regression model with 46 variables is robust as the root mean square error is similar for the train and test sample
* The prediction error is low when the property value is less than 2M USD and prediction error is high for property value greater than 2M USD
* BDSP, RMSP, ELEP, GASP, CONP, SMP, SMOCP, MRGP have the highest influence because of the high coefficient values
* Further enhancements are needed and for better interpretation and prediction accuracy:
  + To explain the impact of the variables on the property value it would needed to create dummies for factors
  + Low root mean square error is possible
    - if we first cluster the data and have different models for the different clusters
    - oversample houses where property price is greater than 2M or under sample houses where property prices is less than 2M